turning knowledge into practice

Research Triangle Park, North Carolina

Disclosure-treated Surveillance Data for Developing Enhanced Detection Tools for High Risk Profiles

A.C. Singh[1], F. Yu[1], D.H. Wilson[1], and J. D. Eyerman[2]

[1]Statistics Research Division, [2]Survey Research Division

RTI International

Presentation to PHIN,
Atlanta, GA, May 27, 2004

Outline

- Relevance of Public Use Files (PUFs) to PHIN Goals
- PUFs and Surveillance Data
- Example of Public Use Surveillance Data
- PUFs: a Preamble
- Inside vs Outside Intrusion Scenarios
- Solutions
  - Deterministic vs. stochastic selection of records for treatment
  - RTI's solution – MASSC
- Example
- Summary

Relevance of PUFs to PHIN Goals

- Goal: *"PHIN will enable consistent exchange of response, health, and disease tracking data between public health partners…"* (http://www.cdc.gov/phin/index.htm).
- The purpose of this goal, under the PHIN-identified industry standard for Analysis and Visualization, in particular, is to be able to use innovative surveillance tools for detecting public health events due to high risk- profile individuals.

PUFs and Surveillance Data

- Important to have PUFs with surveillance data available to researchers at large within an agency as well as available to **public health partners**:
  - Academic research community
  - Industry partners
  - Policy makers
  - Other federal departments and agencies
- These public health partners require PUFs for a broad group of researchers at the tool development stage and, later, the selected tool can be tested on the original data under a secure environment. This has implications on Research and development, Business planning, Evidence based policy, and Reduction of data collection burden on populations.

Examples of Public Health Surveillance Data

- Surveillance systems operated by NCID (http://www.cdc.gov/ncidod/osr/site/surv_resources/surv_sys.htm)
  - 30 listed on the website
  - None had PUFs (downloadable, analyzable, data matrices)
  - Some report tables that **may be** vulnerable to an inside intruder
- CDC Wonder (http://wonder.cdc.gov/)
  - WONDER provides a single point of access to a wide variety of reports and numeric public health data
  - 50 data sources listed
  - 24 provide numerical data to query or download

PUFs: a Preamble

- Since the formal enactment of HIPAA regulations (1996), research on data disclosure has become a very active area.
- RTI's statisticians got involved five years ago through their work on the Drug Survey (NSDUH).
- RTI needed to address the difficult problem of inside intrusion for NSDUH; intruder knows the presence of his target in the database, e.g., father would like to know his son's drug behavior.
- Public Use Files (PUFs)

Inside vs. Outside Intrusion Scenarios

- "Disclosure by response knowledge" – an important inside intrusion scenario from respondent's perspective.
  - **IVs:e.g.,** for BRFSS data, age group, race, gender, education, income, height, weight, freq of eating fruits, flu shots, etc.
  - **SVs:** e.g., for BRFSS data, asthma condition, diabetes condition, # permanent teeth removed, drinking alcohol and driving car, reason for HIV test, method of birth control, etc.
- A respondent identifies his own record and is concerned about its disclosure by someone who might know enough about him to identify his record.
- Protecting against inside intrusion automatically protects against outside intrusion; provides an upper bound on disclosure risk.

Disclosure Risk from an Inside Intruder

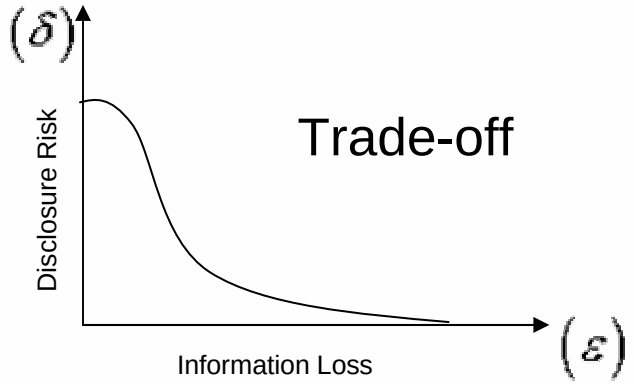
- consider a hypothetical data with 10 observations.

(IVs= age, gender; SV=positive Hypertension diagnosis)

Raw Data Before Treatment					
Obs	Age	Gender	Diag	Status before treatment	Risk Status
1	4	F	N	Nonunique double	Not at Risk
2	2	F	Y	Nonunique double	At Risk
3	2	F	Y	Nonunique double	At Risk
4	1	M	Y	Unique	At Risk
5	4	F	N	Nonunique double	Not at Risk
6	1	F	Y	Unique	At Risk
7	3	M	N	Nonunique triple	Not At Risk
8	2	M	Y	Unique	At Risk
9	3	M	Y	Nonunique triple	Not At Risk
10	3	M	Y	Nonunique triple	Not At Risk

Need for Disclosure Treatment and Control on Information Loss (a conundrum)

- Some treatment of perturbation and suppression needed to protect from disclosure.
- Any disclosure treatment leads to information loss.
- How to balance the tension between disclosure risk (δ) due to limited amount of perturbation and suppression, and information loss (ε) due to introduction of bias and variance?



- Useful to have (ε, δ) measures for any process of disclosure treatment.

Deterministic vs. Stochastic Framework

Deterministic Selection for Treatment: Only records at risk are treated; the risk goes to zero but it may lead to high information loss. Also, there is no protection against new IVs.

Stochastic Selection for Treatment: All records are subject to treatment but only a small random subset is actually treated; leads to low information loss and protection against new IVs. However, risk is not zero but small after treatment.

Note: Need a probabilistic/stochastic framework to measure and control disclosure risk and information loss.

Concept behind RTI's MASSC

- Under inside intrusion, a database is the finite population.
- Subtle analogy between census taking (has high monetary cost) and releasing the original database (has high disclosure cost).
- Take a well-designed sample from the finite population/database:
  - Stratify for over/under sampling //create risk strata for over/under treatment)
  - Impute for item nonresponse // perturb at random
  - Sample selection // random non-suppression
  - Weight calibration to reduce bias due to nonresponse and variance due to sampling//same

Process of MASSC
(A nonsynthetic approach)

Steps:

I: Micro Agglomeration
(for creating risk strata to check & control the number of records at risk of disclosure)

II: Optimal Random Substitution
(to introduce uncertainty primarily about the identity of a target)

III: Optimal Random Subsampling
(to introduce uncertainty primarily about the presence of a target.)

IV: Optimal Calibration
(to reduce bias due to substitution and variance due to subsampling.)

A Simple Illustrative Example (Micro Agglomeration)

Raw Data			
Obs	Age	Gender	Diag
1	4	F	N
2	2	F	Y
3	2	F	Y
4	1	M	Y
5	4	F	N
6	1	F	Y
7	3	M	N
8	2	M	Y
9	3	M	Y
10	3	M	Y

Data After Micro Agglomeration				
Obs	Age	Gender	Diag	Status before treatment
4	1	M	Y	Unique; at risk
6	1	F	Y	Unique; at risk
8	2	M	Y	Unique; at risk
2	2	F	Y	Nonunique double; at risk
3	2	F	Y	Nonunique double; at risk
1	4	F	N	Nonunique double; not at risk
5	4	F	N	Nonunique double; not at risk
9	3	M	Y	Nonunique triple; not at risk
7	3	M	N	Nonunique triple; not at risk
10	3	M	Y	Nonunique triple; not at risk

Note: Under Inside Intrusion, unique records with sensitive values are at risk, and nonunique records with common sensitive values of a SV are at risk.

A Simple Illustrative Example (Substitution)

Data After Micro Agglomeration					After Substitution			
Obs	Age	Gender	Diag	Status before treatment	Age	Gender	Diag	
4	1	M	Y	Unique; at risk	1	M	Y	
6	1	F	Y	Unique; at risk	1	M	Y	
8	2	M	Y	Unique; at risk	2	M	Y	
2	2	F	Y	Nonunique double; at risk	2	F	Y	
3	2	F	Y	Nonunique double; at risk	2	F	Y	
1	4	F	N	Nonunique double; not at risk	4	F	N	
5	4	F	N	Nonunique double; not at risk	3	M	N	
9	3	M	Y	Nonunique triple; not at risk	3	M	Y	
7	3	M	N	Nonunique triple; not at risk	3	M	N	
10	3	M	Y	Nonunique triple;not at risk	2	M	Y	

A Simple Illustrative Example (Subsampling)

Data After Micro Agglomeration					After Substitution			After Subsampling
Obs	Age	Gender	Diag	Status before treatment	Age	Gender	Diag	Status after treatment
4	1	M	Y	Unique; at risk	1	M	Y	Sampled out
6	1	F	Y	Unique; at risk	1	M	Y	Pseudo-unique
8	2	M	Y	Unique; at risk	2	M	Y	Pseudo-nonunique double
2	2	F	Y	Nonunique double; at risk	2	F	Y	Pseudo-unique
3	2	F	Y	Nonunique double; at risk	2	F	Y	Sampled out
1	4	F	N	Nonunique double; not at risk	4	F	N	Pseudo-unique
5	4	F	N	Nonunique double; not at risk	3	M	N	Pseudo-nonunique triple
9	3	M	Y	Nonunique triple; not at risk	3	M	Y	Nonunique triple
7	3	M	N	Nonunique triple; not at risk	3	M	N	Nonunique triple
10	3	M	Y	Nonunique triple;not at risk	2	M	Y	Pseudo-nonunique double

Confidentiality Diagnostics

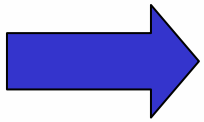
$$\delta_u, \delta_{nu(d)}, \delta_{nu(t)}, \delta_{nu(o)}$$

A Simple Illustrative Example (Calibration)

Data After Micro Agglomeration					After Substitution			After Subsampling	After Calibration
Obs	Age	Gender	Diag	Wt	Age	Gender	Diag	Status after treatment	Wt
4	1	M	Y	1	1	M	Y	Sampled out	0.00
6	1	F	Y	1	1	M	Y	Pseudo-unique	0.83
8	2	M	Y	1	2	M	Y	Pseudo-nonunique double	0.83
2	2	F	Y	1	2	F	Y	Pseudo-unique	2.50
3	2	F	Y	1	2	F	Y	Sampled out	0.00
1	4	F	N	1	4	F	N	Pseudo-unique	2.50
5	4	F	N	1	3	M	N	Pseudo-nonunique triple	0.83
9	3	M	Y	1	3	M	Y	Pseudo-nonunique triple	0.83
7	3	M	N	1	3	M	N	Pseudo-nonunique triple	0.83
10	3	M	Y	1	2	M	Y	Pseudo-nonunique double	0.83

A Simple Illustrative Example (MASSC Result)

Raw Data			
Obs	Age	Gender	Diag
1	4	F	N
2	2	F	Y
3	2	F	Y
4	1	M	Y
5	4	F	N
6	1	F	Y
7	3	M	N
8	2	M	Y
9	3	M	Y
10	3	M	Y



Data After MASSC				
Obs	Age	Gender	Diag	Wt
6	1	M	Y	0.83
8	2	M	Y	0.83
2	2	F	Y	2.50
1	4	F	N	2.50
5	3	M	N	0.83
9	3	M	Y	0.83
7	3	M	N	0.83
10	2	M	Y	0.83

Quality Diagnostics

- Average absolute relative bias
- Average decrease in precision

Summary

MASSC Value vs. Alternatives (IV = Identifying/Intrusion Variable)

Confidentiality/Quality	Alternative Methods (Deterministic Framework)	MASSC (Stochastic Framework)
Data Confidentiality		
Selected IV's	High Protection (low risk)	High Protection (low risk)
Other IV's	No protection (maximum risk)	Moderate to high protection (low to moderate risk)
Measure of Disclosure Risk	Not Generally Available	Available
Data Quality		
Bias	High	Low
Variance	Zero (typically)	Low to Moderate
Information on IV's	Coarse Categories	Fine Categories
Standard Analysis Tools	Not Applicable	Applicable
Measure of Info Loss	Not Available	Available

Concluding Remarks

- MASSC is currently being applied to NHIS and NHANES datasets. It can be used for BRFSS dataset and other CDC datasets.
- MASSC treatment for very large datasets can be made computationally feasible by partitioning into smaller subsets and treating each subset separately.
- For MASSC application, a dataset is required to be complete with respect to key IVs, SVs, and AVs. Therefore, some imputation may be needed. There is, of course, less information with imputed data than complete data, but imputation also provides additional protection from intrusion.
- Standard survey data analysis software is applicable to MASSC-treated dataset.

Concluding Remarks (Cont.)

- Highly sensitive data not previously available to researchers can be made available after MASSC treatment.
- In data mining applications for detecting rare events or characteristics of small subgroups, MASSC-treated surrogate data can be used by researchers at large, and then the final analysis on the original dataset can be performed under tight security.
- Periodic updating of databases collected longitudinally (i.e., more fields over time) can be done by using substitution and subsampling rates used for the initial MASSC as long as disclosure risk remains at a reasonable level.
- Similarly, periodic updating of time series of databases (i.e., more records over time) can be done as long as information loss remains at a reasonable level.

Contact for Additional Information

Dr. Michael Samuhel, Director

Research and Development
RTI International
919-541-5803
samuhel@rti.org